100% ACTUAIRES & DATA SCIENCE
Jeudi 17 novembre 2022

Institut des ACTUAIRES

Méthodes d'extrapolation et d'interpolation - Applications à la gestion des risques

Olivier BRU

Senior Manager

Shohruh MIRYUSUPOV, PhD

Research Advisor

Idriss TCHAPDA, PhD

Actuaire,
Institut des Actuaires

Qui sommes-nous en quelques chiffres ?

2011

Depuis 11 ans, **MPG Partners** se développe de manière organique

1 secteur

MPG Partners est un cabinet de conseil en stratégie opérationnelle et management exclusivement dédié aux Institutions Financières
(Banques de détail, BFI, Etablissements financiers spécialisés, Banques Privées, Gestionnaires d'actifs, Assurances)

3D une offre en 3 dimensions

- Conseil métier et règlementaire
- Conseil augmenté par l'IA
- Conseil en transformation

1 Laboratoire de R&D

Un programme de R&D lancé il y a 10 ans avec le **CNRS** qui a donné naissance à **CoDPy**, librairie de **Machine Learning** accessible sous python qui permet de résoudre les problèmes les plus complexes en ingénierie financière

50 collaborateurs

1 volonté

Apporter de la **valeur Scientifique** aux projets opérationnels

Top 500
des entreprises performantes :
DÉCIDEURS MAGAZINE

Labellisée
Fintech Innovante :
Finance Innovation
Projet labellisé

Membre du réseau
BPI EXCELLENCE :
Entreprise membre de bpifrance EXCELLENCE 2018

TOP 50
des partenaires d'affaires :
PLIMSOLL
VOTRE BUSINESS INTELLIGENCE

CoDPy, fruit de la R&D

20 publications

qui portent sur les fondements mathématiques de CoDPy et les algorithmes (les méthodes a noyau)

3 revues Alpha

WILMOTT MAGAZINE

COMPTE RENDU POUR L'ACADEMIE DES SCIENCES

COMPUTER METHODS IN APPLIED MECHANICS AND ENGINEERING

5 partenaires clés

Entreprise membre de bpifrance EXCELLENCE 2018

cnrs

Finance Innovation
Projet labellisé

Bloomberg

TSE
Toulouse School of Economics
École d'économie de Toulouse

10 années de R&D

Financées par MPG Partners et en étroite collaboration avec le **Laboratoire Jacques Louis Lions du CNRS**

7 use cases clients

GROUPE BPCE
Coopératifs, banquiers et assureurs autrement.

Nordea

SOCIETE GENERALE

CAISSE D'EPARGNE

Motivations

- Existe-t-il une méthode pour interpoler et extrapoler les données lorsque la distribution est
  - non gaussienne
  - non linéaire
  - en haute dimension

- Peut-on avoir un modèle génératif interprétable qui calibre une distribution inconnue ?

L'interpolation et l'extrapolation avec CodPy

- Qu'est-ce que **CodPy**?
  - **CodPy** est une bibliothèque open-source d'apprentissage automatique basée sur la théorie de l'espace de Hilbert à noyau reproduisant (RKHS) et le transport optimal.
- Pourquoi **CodPy**?
  - Modèles d'apprentissage automatique interprétables
  - Erreurs de prédiction par les fonctions de la discrépance
  - Entraînement rapide du modèle
  - Entraînement avec un petit nombre d'exemples
  - Les méthodes d'apprentissage automatique supervisées, non supervisées et semi-supervisées.
  - Calcul des sensibilités et solution des EDP (opérateurs différentiels à noyau)

Contexte

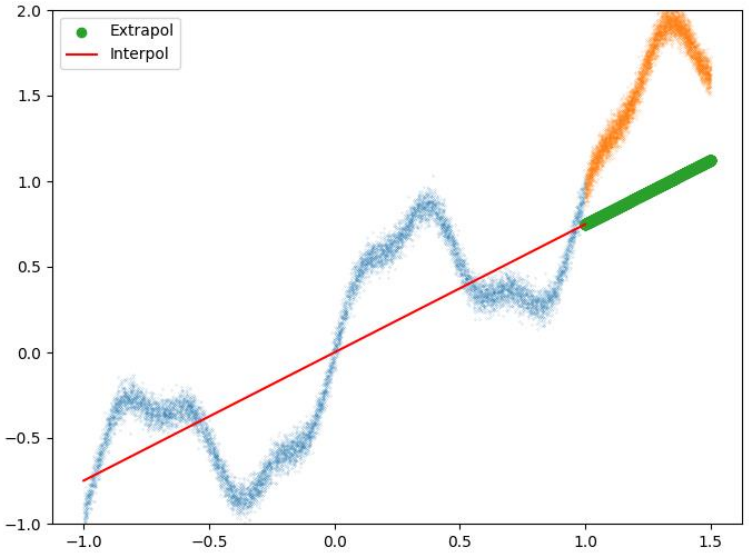
- Étant donné un ensemble de variables explicatives $X \in \mathbb{R}^{N_x, D}$ et de variables de réponse $f(X) \in \mathbb{R}^{N_x, D_f}$, nous voulons établir une relation entre elles, ce processus est défini comme l'apprentissage automatique supervisé.

- La pratique courante consiste à diviser les données en le train set $\{X, f(X)\}$ et le test set $\{Z, f(Z)\}$. Le processus consiste à trouver une relation entre les variables X et $f(X)$, puis à généraliser le modèle à des variables inconnues, c'est-à-dire l'ensemble de test.

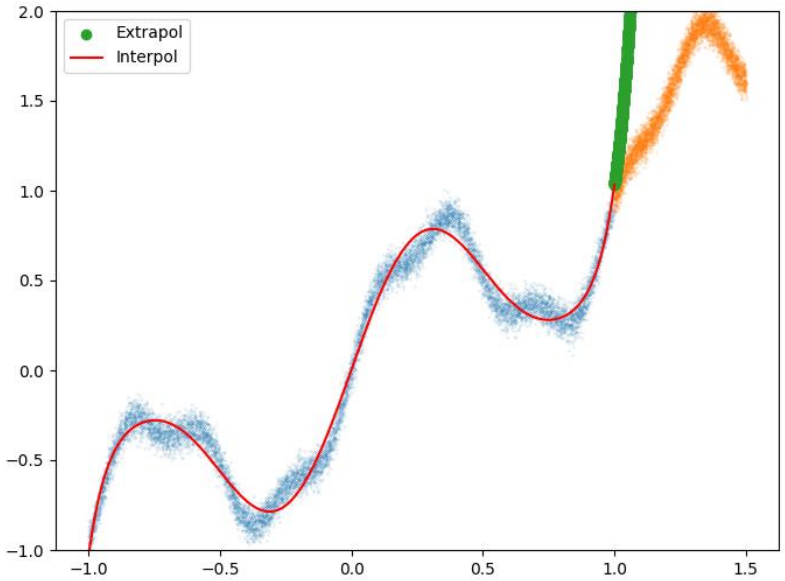
- La première étape est l'interpolation, et la seconde l'extrapolation.

Exemple: Ajustement de courbe et la prévision

- Les ensembles $\{X, f(X)\}, \{Z, f(Z)\}$ correspondent aux données que l'on veut interpoler et extrapoler respectivement.



La régression linaire



La régression polynomial

Quelles méthodes peut-on utiliser pour les problèmes d'interpolation et d'extrapolation ?

- Modèles de régression linéaire et polynomiale
- Réseaux de neurones
- XGBoost
- Méthodes à noyaux:
  - Modèles de régression linaire et non linéaires
  - Les SVMs
  - Les processus gaussiens

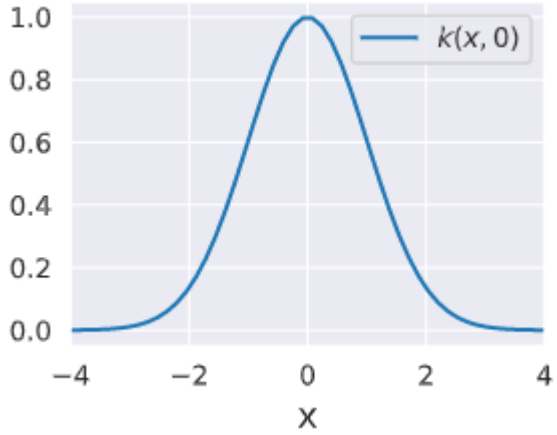
RKHS: la définition

- Nous appelons une fonction $k: \mathcal{X} \times \mathcal{X} \mapsto R$ un noyau si elle est symétrique et définie-positive.
- L'espace de Hilbert à noyau reproduisant (RKHS) $\mathcal{H}_k$ est un espace de Hilbert, généré par le noyau k, dont le produit scalaire satisfait à la propriété de reproduction suivante

$$k(x, y) = \langle k(x, \cdot), k(y, \cdot) \rangle_{\mathcal{H}_k}, \forall (x, y) \in \mathcal{X} \times \mathcal{X}$$

- Exemple: le noyau Gaussien

$$k(x, y) = e^{-\frac{(x-y)^2}{2\sigma}}$$



RKHS: l'interpolation – l'extrapolation

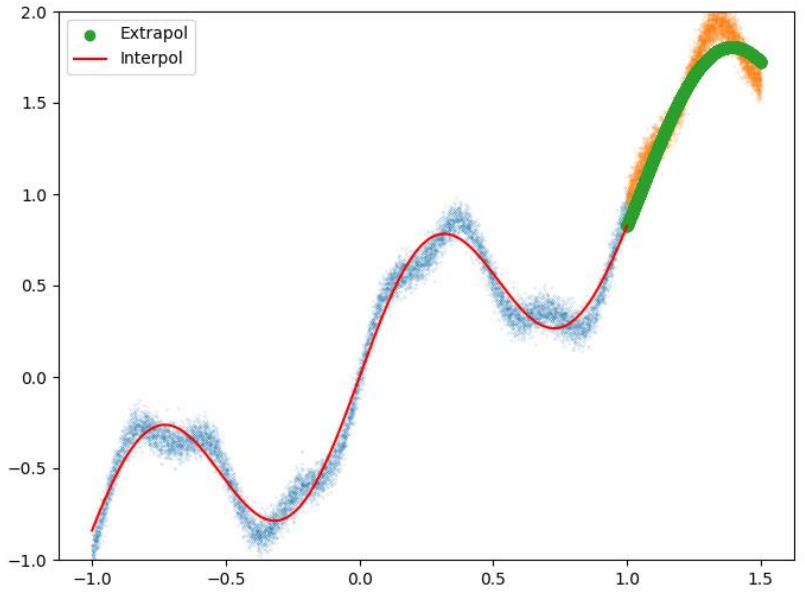
Considérons la matrice de noyau $K(X,Z) := \left(k(x^n, z^m)\right)_{n,m=1}^{N_x,N_z}$ comme une matrice définie positive induite par le noyau Gaussien.

L'interpolation/extrapolation basée sur le noyau est donnée par

L'interpolation: $f_X = \mathcal{P}_k(X,X)f(X) = K(X,X)K(X,X)^{-1}f(X)$

L'extrapolation: $f_Z = \mathcal{P}_k(X,Z)f(X) = K(Z,X)K(X,X)^{-1}f(X)$

La complexité algorithmique est de l'ordre de $\mathcal{O}(N_x^3)$.
Coûteux pour certains problèmes



Peut-on traiter les grandes bases de données ?

- L'opérateur de projection permet d'interpoler et d'extrapoler efficacement sur un grand nombre d'exemples.

L'interpolation: $\mathcal{P}_k(X,Y,X) \coloneqq K(Y,X)(K(Y,X)K(X,Y)+\epsilon I_d)^{-1}K(Y,X)\,f(X),$

L'extrapolation: $\mathcal{P}_k(X,Y,Z) \coloneqq K(Y,Z)(K(Y,X)K(X,Y)+\epsilon I_d)^{-1}K(Y,X)\,f(X),$

Y est le paramètre interne, $\epsilon \geq 0$ est un terme de régularisation.

- La complexité algorithmique est de l'ordre de

$$D\left((N_y)^3 + (N_y)^2 N_x + (N_y)^2 N_z\right),$$

linéaire dans l'interpolation et l'extrapolation

La discrépance et l'erreur maximale de prédiction

- La discrépance entre $X \in \mathbb{R}^{N_x,D}$ et $Y \in \mathbb{R}^{N_y,D}$ est définie comme

$$D_k(X,Y)^2 := \frac{1}{N_x^2}\sum_{n,m=1}^{N_x} k(x^n,x^m) + \frac{1}{N_y^2}\sum_{n,m}^{N_y} k(y^n,y^m) - \frac{2}{N_x N_y}\sum_{n,m=1}^{N_x,N_y} k(x^n,y^m)$$

- Nous pouvons choisir le noyau k pour contrôler l'erreur de prédiction:

$$|f(Z) - f_Z|_{\ell^2} \leq (D_k(X,Y) + D_k(Y,Z))\ \|f\|_{\mathcal{H}_k}$$

Les cartes

Dans de nombreux cas traitant des données brutes, nous souhaitons utiliser une transformation. La notion de carte S: $k(x,y) \rightarrow (k \circ S)(x,y)$ permet de trouver la transformation adaptée au noyau et au modèle.

Quelques exemples:

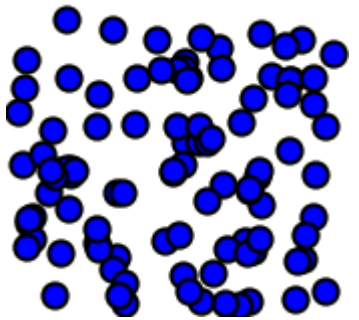
- log-return
- à l'échelle de la distance moyenne
- …

Nous pouvons non seulement générer de nouveaux noyaux en utilisant ces cartes, mais aussi représenter un réseau de neurones

Modèle génératif à noyau basé sur le transport optimal

Étant donné une distribution inconnue $\mathbb{X}$, nous aimerions être en mesure d'en tirer un échantillon. Quelles méthodes pouvons-nous utiliser ?

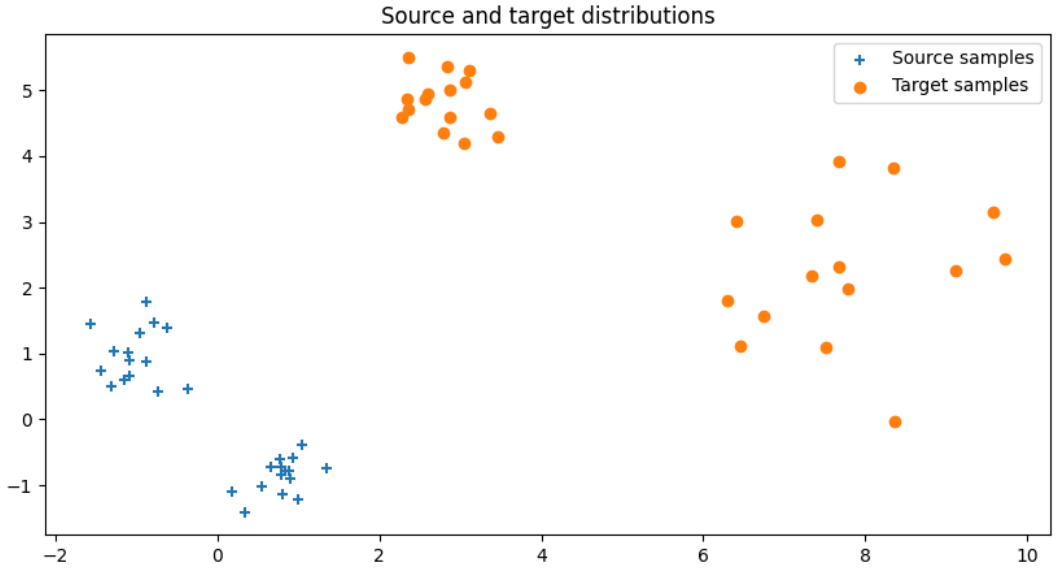
- Méthode du maximum de vraisemblance
- MCMC et méthodes bayésiennes
- Les GANs
- Modèle génératif à noyau basé sur le transport optimal

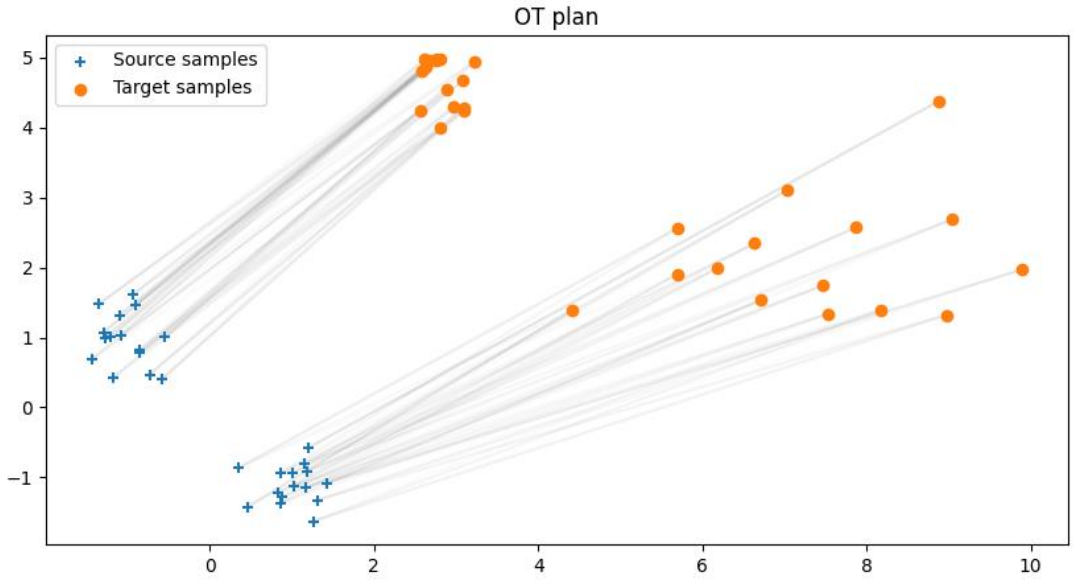


Transport optimal

Nous voulons calculer une carte T transportant de la distribution source $Y \sim \mathbb{Y} \in \mathbb{R}^{N_y, D}$ à cible $X \sim \mathbb{X} \in \mathbb{R}^{N_x, D}$.
T est définie par toute carte de permutation $\sigma : \{1, \dots, N_x\} \mapsto \{1, \dots, N_x\}$

$$T(Y) \coloneqq X^{\sigma} \coloneqq \{x^{\sigma(n)}\}_{n=1}^{N_x}$$



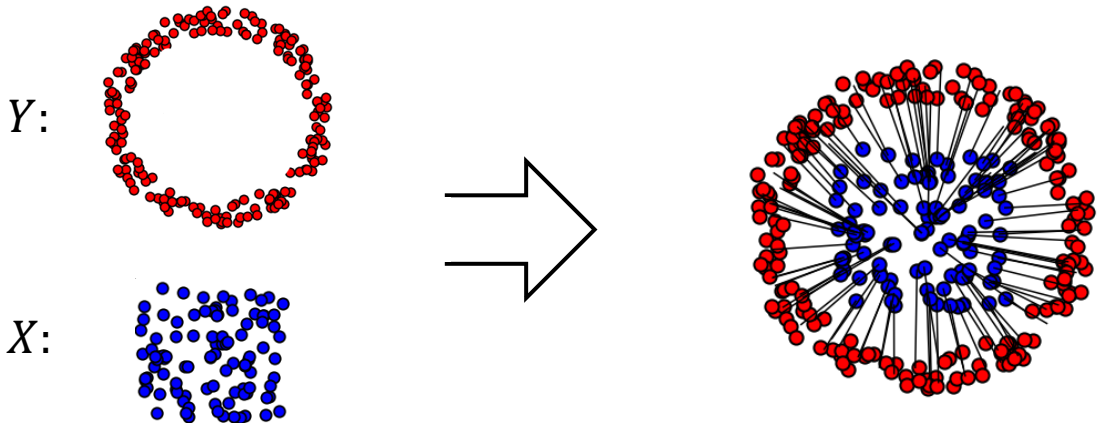


Le plan de transport optimal

- Le plan de transport optimal peut être trouvé en résolvant le problème suivant :

$$\bar{\sigma} = \arg\inf_{\sigma \in \Sigma} Tr\left(M_k(X^\sigma, Y)\right),$$

M_k est la matrice de coût induite par le noyau k.



L'operateur de sampling

- En utilisant les méthodes du noyau, nous pouvons apprendre cette carte de transport.
- Le modèle génératif à noyau basé sur le transport optimal est une méthode d'estimation non paramétrée de la distribution.
- Le modèle génératif est défini comme

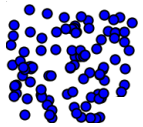
L'operateur de sampling

$$G(z):\ z: \mapsto \mathcal{P}_k(Y, Y, z)\ X^{\overline{\sigma}}, \quad Y \sim \mathbb{Y},\ X \sim \mathbb{X},$$

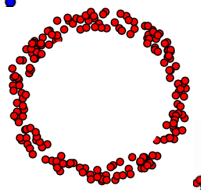
$$G(z) \sim \mathbb{X},\ z \sim \mathbb{Y}$$

Modèle génératif à noyau basé sur le transport optimal

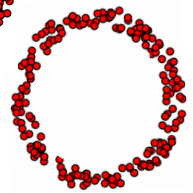
Préparation



Distribution inconnue: $X \sim \mathbb{X}$

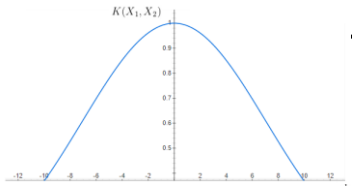


IID: $Y \sim \mathbb{Y}$



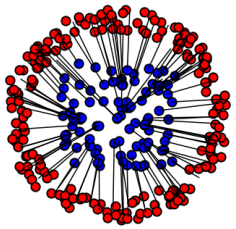
IID: $z \sim \mathbb{Y}$

Le noyau: k



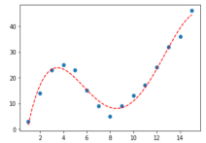
Transport optimal

Calcul de la carte de transport: $T(Y) = X^{\sigma}$



Apprentissage

Calibration du modèle à noyau $\mathcal{P}_k(Y, Y, z)$



Modèle génératif

$G(z)$: $z \mapsto \mathcal{P}_k(Y, Y, z)\, X^{\sigma}$

- L'échantillon statistiquement proche de la distribution inconnue $\mathbb{X}$

CodPy: Interface python simple

- Pour extrapoler, nous utilisons l'opérateur de projection :

```
f_z = op.projection(x = x,y = x, z = z, fx = fx, **kwargs)
```

  - x, fx, z = x correspondent aux données que l'on veut interpoler
  - x, fx, z ≠ x correspondent aux données que l'on veut extrapoler

- Deux hyperparamètres: le noyau et la carte

```
kwargs =  {
'set_codpy_kernel':kernel_setters.kernel_helper(kernel_setters.set_kernel, 0, 1e-8, map_setters.set_map)
}
```

CodPy: Interface python simple

- L'opérateur de sampling :

```
alg.sampler(X, Y, z, M,**kwargs)
```

  - X correspond à la distribution inconnue $\mathbb{X}$.
  - z, Y correspondent aux variables aléatoires IID de la distribution source $\mathbb{Y}$.
  - M est le nombre de variables aléatoires que nous échantillonnons.

- Deux hyperparamètres: le noyau et la carte

```
kwargs =  {
'set_codpy_kernel':kernel_setters.kernel_helper(kernel_setters.set_kernel, 0, 1e-8,
map_setters.set_map)
}
```

Use-case: Génération de données synthétiques

- Nous observons le prix d'un actif $X = (X_1, \dots, X_T)$.
- L'hypothèse markovienne:

$$X_t = X_s e^{(t-s)\mu + \sqrt{t-s}\mathbb{X}}$$

- En utilisant l'opérateur de projection et le transport optimal, comment pouvons-nous générer différents scénarios de marché pour le prix de l'actif ?

- Ex. Prix Google du 01.01.2016 au 31.12.2020,

  T = 1306.

Solution: modèle à noyau et le transport optimal

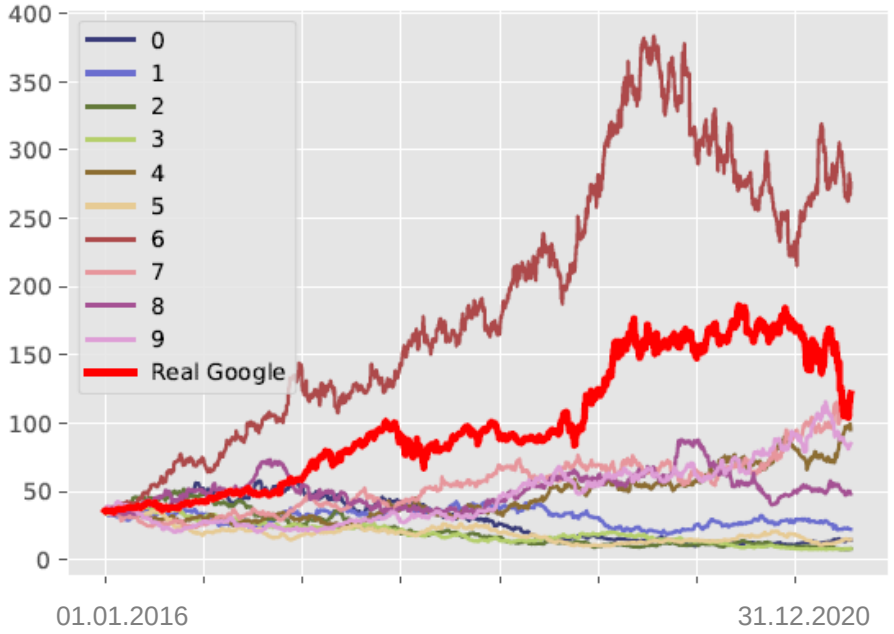
1. La sélection d'un noyau k, l'application d'une carte de log-return:

$$\{\tilde{X}\}_{j=1}^{T} := \{S(X)\}_{j=1}^{T} = \left\{\frac{ln(X_{j+1}) - ln(X_j)}{\sqrt{t^{j+1} - t^j}}\right\}_{j=1}^{T}, \quad \tilde{X} \sim \mathbb{X}.$$

2. L'application de l'opérateur de sampling pour obtenir des échantillons Z :

$$G(z): \ z: \mapsto \mathcal{P}_k(Y, Y, z)\tilde{X}^{\bar{\sigma}}$$

3. L'application de la transformation inverse S^{-1} pour obtenir des scénarios du prix de l'actif.



Use-case : Optimisation du recouvrement de créance (NPL)

Objectif et problématique

- Amélioration opérationnelle : Limiter le recours au recouvrement (ARA) pour les clients devant être traités en Agence
- Expérience client : Limiter l'attrition des clients
- Fraude : Identifier les cas à céder dès l'incident
- Provisions : limiter les impacts financiers pour la Banque

Prédiction de séries temporelles (exposition client)

Calcul de probabilités conditionnelles (défaut client)

Multitude sources de risques (100 variables explicatives)

Approche

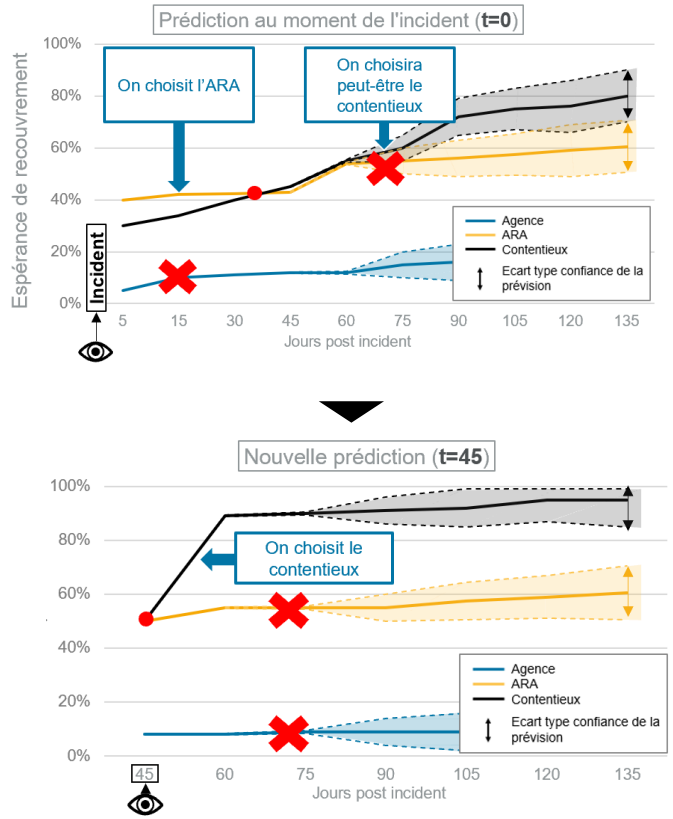
- Construction d'un modèle permettant de prédire la trajectoire du taux de recouvrement sur 300 jours pour chaque étape du processus
  - Agence
  - Recouvrement amiable (ARA)
  - Contentieux
- Constitution d'une base de **100 000 incidents**
- Transformation des variables en séries temporelles
- Analyse de 60 variables et sélection automatique de 7 variables discriminantes pour un modèle de prédiction optimal
- Construction d'un modèle de probabilité de pertes pour l'identification des cessions à réaliser en amont

Résultats

- **90%** de fiabilité obtenue pour le modèle de trajectoires prédictives sur le comportement de paiement des clients
- **98%** de fiabilité obtenue sur le modèle de probabilité de pertes
- Déploiement pilote (MVP) pour une phase de test en **situation réelle** CAISSE D'EPARGNE

Illustration

Les scores obtenus permettent de choisir le traitement opérationnel optimal à chaque pas de temps



Conclusion

Les méthodes à noyaux:

- permettent d'interpoler et d'extrapoler efficacement les données
- n'ont que deux hyperparamètres à choisir;
- permettent de contrôler l'erreur d'extrapolation et sont interprétables;
- ont un lien naturel avec le transport optimal;
- ne nécessitent pas un échantillon de grande taille.

Avec les fonctions CodPy:

- vous pouvez calibrer un processus stochastique markovien;
- vous pouvez traiter des bases de données relativement grandes (CEBPL);
- vous pouvez extrapoler et interpoler avec plus de 10 noyaux différents et cartes.

pip install codpy

Pour aller plus loin

- Clustering
  - Détection des fraudes
- Évaluation d'option et sensibilités
- L'explication PnL
- …

Références

- Ph. G. LeFloch, J.M. Mercier, S. Miryusupov. CodPy: A Python Library for Machine Learning, Mathematical Finance, and Statistics
  - https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3766451
- A. Berlinet and C. Thomas-Agnan, Reproducing kernel Hilbert spaces in probability and statistics, Springer Science, Business Media, LLC, 2004.

Nos Publications Scientifiques

Nos publications scientifiques se découpent en deux volets : une partie théorique qui porte sur les fondements mathématiques de nos méthodes innovantes et une partie algorithmique qui porte sur notre librairie d'algorithmes (Kernel Methods) :

Théorie

- *"A class of mesh-free algorithms for mathematical finance, machine learning and fluid dynamics* – Ce papier est l'essence même de notre approche ;
- *"The Transport-based Mesh-free Method (TMM) and its applications in finance: a review"* – publié dans Wilmott magazine, Ce papier est une presentation générale de notre approche ;
- *"Mesh-free error integration in arbitrary dimensions: A numerical study of discrepancy functions"* – publié dans Computer Methods in Applied Mechanics and Engineering. Ce papier se concentre sur l'erreur d'analyse pour la méthode Reproducing Kernel Hilbert Space (RKHS) ;
- *"A new method for solving Kolmogorov equations in mathematical finance"* - Comptes Rendus Mathématique, Volume 355, 6ème édition, Juin 2017, Pages 680-686. Ce papier explique la stratégie des équations différentielles partielles pour les mathématiques financières et propose un certain nombre d'exemples numériques ;
- *"Revisiting the method of characteristics via a convex Hull algorithm"* – publié dans le Journal of Computational Physics, Octobre 2015 (lien référence) Ce papier traite de l'application de cette méthode pour les lois de conservation ;
- *"A high dimensional framework for financial instruments valuation"* – publié en 2013 ce papier est un premier essai pour décrire notre approche et la résolution d'équations aux dérivées partielles multidimensionnelle en mathématiques financières ;
- *"Optimally transported schemes"* – Papier portant sur le traitement pour le cas à une seule dimension ;

Algorithmes

- *"CoDPY -Tutorial"* – une première introduction de notre librairie, notamment sur le focus du Machine Learning ;
- *"CoDPY - Advanced Tutorial"* – une description technique de notre librairie CoDPY ;
- *"A kernel based reordering algorithm"* – description de l'algorithme central de reorganization de notre approche ;
- *"A kernel based polar factorization and the sampling algorithm with CoDPY"* – en preparation, description de notre algorithme de calcul de factorisation polaire. Cet algorithme est illustré à travers un outil très pratique permettant de produire des échantillons IDD à partir de n'importe quelles distributions d'inputs ;
- *"A kernel-based algorithm to compute conditional expectations"* – en preparation, un des algorithms les plus important pour les applications en Finance. Nous avons benchmarké l'implémentation de cet algorithme, en comparant notre framework à une approche Classique de Neural Network ;
- *"Hedging Strategies for Net Interest Income and Economic Values of Equity"* – une description d'un prototype utilisant CoDPY, ayant pour but la construction de stratégies sophistiquées sur des sujets d'ALM ;
- *"Kernel methods for stress test and reverse stress test"* – description de notre approche Support Vector Machine avancée aux Stress Tests et Reverse Stress Tests ;

A votre disposition pour échanger :

Olivier BRU
Senior Manager, Head of Machine Learning Advisory
Olivier.BRU@mpg-partners.com

Shohruh MIRYUSUPOV, PhD
Senior Advisor, Team Leader R&D
Shohruh.MIRYUSUPOV@mpg-partners.com

Jean-Marc MERCIER, PhD
Director, Head of R&D
Jean-Marc.MERCIER@mpg-partners.com

Idriss TCHAPDA, PhD
Actuaire, Membre de l'Institut des Actuaires
tchapdid@me.com